Marine Makhatadze

TOWARD LEXICAL BUNDLES IN GEORGIAN LEARNER CORPUS OF ENGLISH WITH IMPLICATIONS FOR PEDAGOGIC LEXICOGRAPHY

Abstract

A primary component of fluent language production is the control of complex lexemes, known as lexical bundles, chunks, or clusters. These frequently recurring sequences of words (e.g., as compared to) function as building blocks of discourse, helping to shape meanings in specific contexts and contributing to our sense of coherence in a text.

This paper is based on the corpus analysis with a combination of quantitative and qualitative criteria. To observe the idiosyncrasies and difficulties non-native writers face in producing accurate texts in English, the author compiled the Georgian Learner Corpus of English (GLEAN). Corpus data comprises argumentative essays, reports, narrative essays, and free composition essays within linguistics and literature, as well as blog posts, diaries and political/apolitical newspaper articles from Georgian learners of English. The aims of this research are as follows: a) to identify which lexical bundles (or lexical phrases) are most common in academic prose produced by Georgian learners of English; b) to classify the functions of the most common 3-word or 4-word lexical bundles used in the GLEAN corpus; and c) to highlight the value of adding the learner corpus data (such as illustrative sentences, usage notes, "help boxes") to learner dictionaries. The research results showed that the most frequent bundles in the Georgian Learner English corpus serve the primary function of participant-oriented bundles, which express attitudes. The final product is a list of 42 lexical bundles that cover academic writing in the GLEAN corpus disciplines (literature, linguistics, press, blog posts, etc.) and the lexicographic implications for further research.

Keywords: *lexical bundles, learner corpus, microstructure*

Even the most unlikely of words is found to have secrets.

- Michael Hoey

1. Introduction

The difficulties non-native writers face in producing accurate, effective expository texts in English have prompted many studies on the lexical elements that constitute well-written prose. The ubiquity of lexical phrases in language production has brought phraseology to the forefront and significantly expanded its scope in language acquisition and meta-lexicographic studies. Multi-word units have been studied under many concepts, including lexical phrases, formulae, routines, fixed expressions, prefabricated patterns, lexical bundles, lexeme clusters or conjuncts, or, in yet other terminologies, bound syntagmas, etc. (Zgusta, 1971). However, terminological diversity shares a common feature in that lexical bundles carry meaning.

Under the impetus of corpus linguistics, it has become increasingly clear that formulaic language is pervasive in authentic language use. Studies have shown that more than 60% of language may be formulaic (Altenberg, 1998; Biber *et al.*, 2000; Erman & Warren, 2000). Because of the complexity of the phrasicon, many learners quit studying the foreign language at their A-levels. As Hausmann (1993:17) notes: "Complex lexemes can only be mastered patiently, diligently, continuously and bee-like... But is there a better life imaginable than humming as you fly from flower to flower, collecting the nectar of the vocabulary and occasionally stinging the one who tries to stop you?".

As Lewis (1993: 93) points out, lexical bundles are still seriously undervalued in ESL teaching. If lexical chunks are such a prevalent language feature, one would expect them to be adequately treated in lexicography and language and translation teaching. To address the research gap, I have compiled a learner corpus for Georgian English (GLEAN) learners and investigated the idiosyncrasies of lexical bundles in academic writing and the press. Thus, the aims of this paper are as follows:

1. To determine which lexical bundles are most frequent in academic prose produced by Georgian learners of English;
2. To classify the functions of the most common 3-word or 4-word lexical bundles used in the GLEAN corpus;

3. To highlight the value of adding learner corpus data (such as illustrative sentences, usage notes and "help boxes") to learner dictionaries. Some lexicographic implications of the findings and suggestions for future research are discussed.

As multi-word units and specifically lexical bundles, are everywhere, one of the primary purposes of compiling learner corpora is to understand the needs of language learners, observe their written or spoken production, and ultimately find ways to guide foreign language learners and facilitate the study process.

2. Previous corpus-based investigations of lexical bundles

Before the advent of learner corpora and corpus tools, the study of multi-word lexical units depended on researchers' intuition and what they perceived to be the most repetitive expressions in the language. According to Sinclair (1991), since corpus analysis is particularly well suited to the study of recurrent multi-word units, the use of large corpora and powerful corpus analysis techniques has led to the discovery of a much more comprehensive range of phraseological units than had been investigated in traditional studies of phraseology. In this sense, there are two fundamental methods of identifying multi-word units, such as lexical bundles, collocations and collostructures in corpora, namely the bottom-up and the top-down methods. In this context, Granger and Paquot's (2008:45) proposal to combine "the best of two worlds" and to reconcile the frequency-based and the phraseological approach is a step in the right direction.

A landmark study of such highly frequent, contiguous word sequences is the extensive study of lexical bundles published as a chapter in the Longman Grammar of Spoken and Written English (Biber *et al.*, 2000). The study is based on analysing multi-million-word corpora representing conversation and academic prose. The authors compare spoken and written university registers and deal with uninterrupted lexical sequences of up to six words.

More recently, further refinements to the lexical bundle approach have been offered by authors such as Hyland (2008a), who developed a functional classification of lexical bundles better suited to written research genres, technical domains such as medicine, economics, etc., and Simpson-Vlach & Ellis (2010), who used a combination of statistical measures and teacher insight to produce a "pedagogically-friendly" list of academic formulae and lexical bundles. Corpus-based research has also shown that these multi-word expressions, which come so naturally to native speakers, are a source of difficulty for non-native users of a language, and some highly frequent

phraseological units produced by language learners have also been described as "lexical teddy bears" (Hasselgren, 1995; Nesselhauf, 2005).

Furthermore, research has also been conducted on lexical bundles in specialized academic texts. Gledhill (2000a), for example, discusses the prevalence of "phraseological accent" in technical writing style, which is demonstrated by the frequent use of formulaic constructions that are uncommon in general English. A study of German academics and journalists revealed that they value a particular type of conventionalized multi-word unit, which the author has named the "second-level discourse marker" (Siepmann, 2004).

From a lexicographic perspective, lexical bundles are still largely absent from dictionaries or are difficult to access. This is due to the challenge of representation, as traditional dictionaries focus on defining individual words and their meanings in isolation. Despite the consensus on the importance of multiword units, there is surprisingly little agreement on their production by non-native speakers and the inclusion of learner data in bilingual, specialized EAP dictionaries.

2.1 Functional taxonomy of lexical bundles

Research on the fundamental nature and structural characteristics of lexical bundles is only possible by studying their functional classifications and how they perform in discourse. Biber, Conrad & Cortes (2004:390) proposed a preliminary categorization grounded in the meanings and purposes of lexical bundles in text. Their framework distinguishes among three primary functions: stance expressions, discourse organizers and referential expressions. This subcategorization is suitable mainly for spoken discourse during classroom teaching rather than written discourse. Thus, due to our research questions, we based our study on O'Flynn's (2022:87) functional taxonomy, which was modified according to Hyland's (2008) dimensions. O'Flynn's (2022:87) primary functions of lexical bundles are divided into three categories: Research-oriented bundles, text-oriented bundles, and participant-oriented bundles. Examining these lexical bundles in textual contexts shows they are essential building blocks of discourse associated with basic communicative functions.

2.2 Treatment of lexical bundles in dictionaries

The lexicographical treatment of phraseology, including lexical bundles, can be expected to differ according to the purpose of the dictionary, whether we are dealing with dictionaries for

encoding or decoding on the one hand and for native speakers or foreign language learners on the other. In general, in monolingual dictionaries, the most typical way of including multiword combinations is as sub-lemmata or as microstructural items. Such decisions made by lexicographers make the lexical bundles more difficult to access, even in electronic dictionaries. More importantly, they are often limited to one translation and/or one example sentence. Compare this to the Dutch dictionary Van Dale Groot Woordenboek Der Nederlandse Taal. The multiword combinations are classified by this dictionary as *uitdrukkingen*(expressions), *zegswijzen* (phrases),*gezegden* (sayings), *spreekwoorden* (proverbs), *vergelijkingen* (comparisons) and *formules* (formulae) and are included independently (Bergenholtz *et al*., 2013). However, dictionaries have always been characterized and dominated by word bias, which has led to a situation where macrostructural selection focuses only on single-unit lexemes and not multi-word units.

Firstly, according to Sinclair (1993:30), dictionaries already have problems with idioms and phrases containing more than one word. There are some lemmatization issues, like where in the dictionary should a phrase *that makes two of us* enter? How can the user be cross-referenced to the broad range of equivalent expressions, such as *"I agree*," "*the same with me*," etc.?

Secondly, to complicate matters, sometimes numerous lexical bundles cannot easily be illustrated within the limited compass of a dictionary. With this in mind, it is little wonder that even the largest of native speaker dictionaries contain hardly any entries on multiword expressions or, for that matter, lexical chunks generally. As Gates (1988:99) puts it, there can be an unconscious feeling for the lexicographer that a dictionary is a book that explains words and that vocabulary items larger than the word is beyond its scope capacity.

A similar picture emerges when examining the four most frequently used monolingual English learner's dictionaries. Despite being otherwise excellent examples of corpus-based lexicography, they almost overlook lexical bundles. While a few rare instances, many of which are also common in spoken English, are included, the Oxford English Dictionary has recorded some of the phrases. As for the English-Georgian Learner's dictionaries do not provide comprehensive coverage of lexical chunks; they include only the parts of common multi-word items that can be recorded as single-word entries.

We agree with Sinclair's (2010:37) opinion that multi-word lexical items should be awarded an independent headword status because they represent a much broader concept than idioms, and we should give them the same status as the usual headwords. Head phrase status may

not be realistic for all types of phraseological units. Still, it is undoubtedly desirable for three or four-word lexical bundles, such as *in accordance with, on the other hand*(Granger 2018:20). One of the main aims of our study is to find some ways of giving lexical bundles proper attention in bilingual English-Georgian dictionaries.

3. Methodology

3.1. Data collection Procedures

The Georgian Learner Corpus of English (GLEAN) was compiled over three years and consists of two core modules: written essays and newspaper articles. The abbreviation of our learner corpus is symbolic, as the verb "to glean" means to gather or pick up the ears of corn after the reapers. This metaphorically reflects our research process, which involves searching for and collecting the peculiarities of word meanings in the language.

This article presents various types of texts according to the fundamental structural criteria. Therefore, the GLEAN corpus comprises almost 10 million words (precisely 9,812,931 tokens) and includes argumentative essays, reports, narrative essays and free composition essays within the disciplines of linguistics and literature as well as blog posts, diaries and political/apolitical newspaper articles from Georgian learners of English. Our learner corpus has been automatically POS-tagged and annotated semantically. Grammatical annotation is an essential feature of corpus texts, making them more valuable. Each word has been automatically assigned a part-of-speech tag or a code giving information about the word class of that particular word. The tool used to tag our learner corpus is CLAWS, a program developed at Lancaster University, with the accuracy rate of the tagging to around 98-99% (Hoffmann *et al.,* 2008). With these tags, it is possible to distinguish instances of *sharp* used as a noun from *sharp* used as a verb, adverb or adjective.

3.2. Research participants

The corpus of Georgian learners of English has been set up to create a tool for lexical research focusing on three or 4-word lexical bundles. The learner corpus consists of the academic essays written by Georgian learners (BA, MA students) of English at Ivane Javakhishvili Tbilisi State University, with a concentration in English Philology. The average age of the participants is 20 (median = 20); 86% are females and 14% - males. The predominance of females is expected, as is the case with all humanities courses. Thus, although the data may appear poorly distributed, it is representative of the population

found in the English Philology degree course. The participants are all Georgian native speakers with Georgian native-speaker parents, and they attended primary and secondary schools with Georgian as the medium of instruction (Makhatadze, 2023). The average number of years studying English at school is 8 (median = 8, IQR = 1), while the average number of years studying English at university is 3 (median = 3).

The majority are in their first term of full-time studies. The length of the essays varies between 120-900 words. The essays were written during the seminars, and the articles were out of class, with a deadline of 1-2 weeks. Consent to participate and permission to use essays in the corpus were given in oral form, and students also completed a questionnaire that provided information for the meta-information. Usually, a learner corpus may contain learner-related variables (age, gender, and other information, e.g. study mobility abroad) and other metadata, such as L1 and parents' L1. These variables are collected because they may provide insights into any potential influences the learner may have been subject to (Granger, 1998a, p. 8).

On the other hand, the average age of the authors of the texts included in the press sub-corpus is 26 years. 72% of them are female, 28% are male. According to the information received from the questionnaire, it is clear that the average number of years studying English is 11 years. The limit of each text included in the press sub-corpus is 500 words. Figure 1 shows the results of the self-evaluation of English proficiency by the participants.

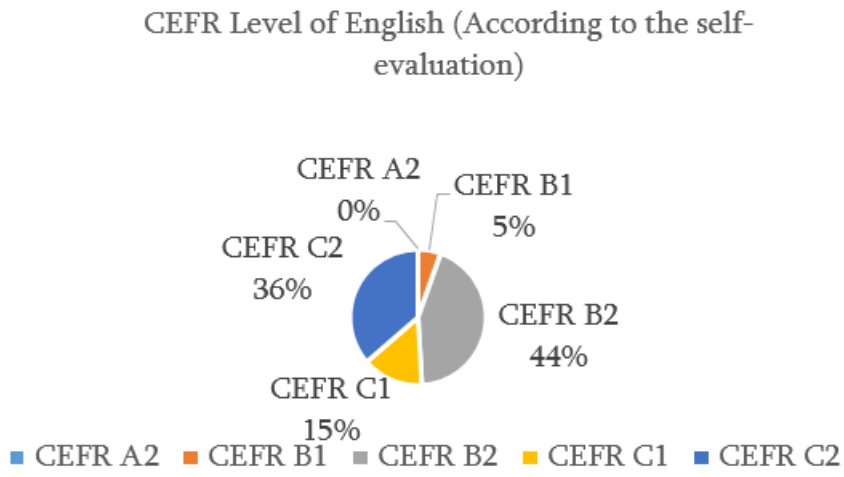

Figure 1 Results of English proficiency (self-evaluation)

3.3. Quantitative analysis of the GLEAN corpus to derive a list of lexical bundles

In our research, the n-gram function of the AntConc (Antony, 2014) was used to derive a list of three- or four-word lexical bundles occurring with a minimum frequency of 10 per million words (PMW). Multi-word units can be investigated using the Word Clusters Tool, which displays clusters of words surrounding a search term and orders them alphabetically or by frequency. An alternative way to search for multi-word units is to find lexical bundles (Biber *et al.*, 2000), equivalent to n-grams, where n can usually vary between two and five words. As our learner corpus has been POS-tagged, we used not only the automatic analysis of the lexical bundles, but we disseminated the chunks by the grammatical category tags; for example, we used *_APPGE* tag for the possessive pronoun and pre-nominal, *_II* for the general preposition and so on.

Several decisions have been made regarding the length and frequency of the lexical bundles in this study. Regarding length, most learner corpus studies focus exclusively on three or four-word bundles because they perform a more comprehensive range of functions than three- or five-word chunks. Table 1 shows the list of lexical bundles in the Georgian Learner Corpus of English.

Table 1 lexical bundles in Georgian Learner Corpus of English

plays an important role	*in recognition of*	*our main goal*
as a result of	*in response to the*	*our thoughts are with*
as for the	*in the case of*	*the context of*
as of today	*in the focus of*	*the need for*
as well as the	*in the middle of*	*to learn more about*
at the beginning of	*in the process of*	*the extent to which*
at the heart of	*in the wake of*	*with the support of*
at the same time	*in this regard*	*also highlighted the fact*
by the end of	*it is important for*	*most common reason*
in a number of	*it is impossible to*	*I believe that the*
in accordance with	*on the backdrop of*	*I remain confident that*
in addition to	*on the contrary*	*I wish to remind*
in an attempt to	*on the one hand*	*no matter how*
in comparison to the	*on the top of*	
in honour of		

3.4. Qualitative analysis of the functions of the bundles in GLEAN corpus

Retrieved lexical bundles were classified qualitatively according to a functional taxonomy, suggested by O'Flynn (2022:87), where the primary functions of lexical bundles are divided into three categories and are well-suited for the written academic prose. The method was qualitative, carried out by the researcher alone, and should not be considered absolute. Some bundles were more difficult to categorize, as the lexical bundles seemed to have multiple functions, such as *at the same time,* however, we categorized them according to which function seemed most dominant based on their use in context. Table 2 shows the above-mentioned functional taxonomy of the lexical bundles.

Table 2 O'Flynn's (2022) functional taxonomy modified

Research-oriented bundles

Function	**Description**
Location	Indicating time/place
Procedure	how or why something is done
Quantification	the quantity or extent of something
Abstract description	an abstract property of something

Text-oriented bundles

Transition signals	Establish additive, comparative, contrastive links between elements.
Resultative signals	Mark inferential or causative relations between elements.
Framing signals	Specifying a context or limiting conditions.

Structuring signal	Specifying to orient the reader.

Participant-oriented bundles

Stance features	A degree of importance, certainty or possibility.
Engagement features	Address the reader directly.

4. Results and discussion

4.1. Frequencies of lexical bundles

The full frequency data for the 42 lexical bundles can be found in Appendix A. The most frequent bundle is *as of today*, with a normalized frequency of 123.658 (see Table 3). After this, the frequencies per million words diminish at a relatively stable rate until the lowest frequency bundle, *also highlighted the fact* (normalized frequency 0.205). Overall, the 42 bundles occur 9,847 times in the 9.8 million word corpus.

Table 3 Raw and normalized frequencies (PMW) of the top ten lexical bundles

Order	**Lexical Bundle**	**GLEAN corpus Raw frequency**	**Frequency (per million words)**
1.	As of today	1,209	123.658
2.	As well as the	1,173	119.976
3.	As a result of	864	88.371
4.	By the end of	794	81.212
5.	In addition to	627	64.131
6.	At the same time	595	60.858
7.	In this regard	390	39.890
8.	In the process of	384	39.276
9.	As for the	383	39.174
10.	In accordance with	383	39.174
..	…	…	…
42.	Also highlighted the fact	2	0.205

The results revealed some lexical errors and they are particularly interesting from the lexicographic perspective which will be discussed in the following sections.

4.2. The functions of lexical bundles

4.2.1. Lexical bundles to help writers present and discuss content

The next step in analysing target bundles is categorising them in terms of their primary discourse-pragmatic functions. O'Flynn's (2022) classification scheme was beneficial for the present study, as it is adapted to the specific concerns of research-focused written genres rather than only for the spoken discourse. However, this framework was treated as a starting point, as it was necessary to make minimal changes to the categories to reflect the functions performed by the lexical bundles more accurately. We analysed the target bundles in their keyword in contexts (KWIC) and determined the specific functions they perform.

To start with, according to the learner corpus data we retrieved from the Georgian learner corpus, research-oriented bundles with the function of location are: *as of today, by the end of, at the beginning of, in the middle of, at the heart of, on the top of, at the same time*. There are seven location bundles, each of which is used almost exclusively for locating events temporally, and some of them indicate time (1) and place (2):

(1) *We cannot deal with everything* ***at the same time****.*

(2) *Unlike other opposition parties, the holding of repeat elections is not* ***on the top of*** *their list of demands.*

The next sub-type is procedure bundles. A particularly interesting example of a procedural bundle is *in the wake of,* found only in political newspaper articles (3).

(3) *Minister also highlighted the Georgian Government's efforts to assist Ukrainian people* ***in the wake of*** *Russian aggression.*

As for the function of quantification, some of these bundles were used abstractly. They did not include precise measurement or counting (4). Quantification of this type reflects the abstract subject matter and is less likely to be found in the scientific genre. The other examples of

quantification bundles are more precise and refer to one countable element of something, but still in most cases, these refer to an intangible entity (5).

(4) *In order to determine* ***the extent to which*** *their financial transactions in 2008 contained elements of money laundering in accordance with international standards and Georgian law.*

(5) *The development of the capital market, named by the Government* ***as one of the*** *most important directions of the country's economic policy.*

The final functional sub-type is abstract description, which includes two bundles *the context of the, the nature of the, on the backdrop of* (6, 7). The fact that there are two bundles serving the primary function of abstract description and no bundles serving the primary function of physical description highlights the focus on abstract constructs in essay writing and the press genre.

(6) *Recently,* ***in the context of the*** *pandemic, we once again felt this support.*

(7) *The paper will focus on exploring* ***the nature of the*** *word polysemy, which features the innate freedom of the language.*

4.2.2 Bundles to help writers organize their text

There are lexical bundles that have the function of helping writers to organize their text. The large number of textual bundles may reflect the more discursive nature of the fields included in the GLEAN corpus. Academic writing in the humanities, especially in literature and linguistics, is, after all, a kind of rhetorical performance (North, 2005), requiring clear and well-organised arguments.

(8) ***The relationship between the*** *two deteriorated for several reasons.*

(9) *The relationship crisis,* ***as well as the*** *lack of trust and love became destructive for the couple.*

(10) ***As for the*** *skip, affirming feedback definitely seems appropriate.*

As for the resultative signals, they mark causative relations between some elements and are found in learner corpus: *as a result of.* The majority of the contexts where this lexical bundle is

found are in the political articles (11). It typically marks the effects caused by a social, political or historical event.

(11) ***As a result of*** *reforms implemented by us, annual expenses for general education have been reduced by 60 per cent for each family.*

Framing signals interpret or explain preceding or forthcoming text. The high number of framing signals may be attributed to the epistemology of the Arts and Humanities disciplines. In these fields, writers have to work harder to persuade their readers (13) (North 2005).

(12) *I am glad that we achieved this goal* ***despite the fact that*** *the situation around us has not improved.*

(13) *Our hopes in this regard are nourished* ***by the fact that*** *Georgia is a land of new opportunities.*

Structural signals are somewhat expected in a corpus comprising BA and MA papers. These lexical bundles help the writer to continue to orient the reader throughout the paragraphs (14). This function is mostly served in the corpus by 3-word bundles: *as discussed above, as noted earlier, etc.*

(14) ***As noted earlier****, the followers of poetry embraced freedom and found it in artistic expression and emotion.*

4.2.3. Bundles to help writers express their attitudes

The most frequent bundles in the Georgian Learner English corpus are the bundles serving the primary function of participant-oriented bundles, which serve the purpose of expressing attitudes. These types of bundles may be used a lot because student essays are a narrative genre in which the writer is expected to present and discuss ideas in a manner which demonstrates opinion. Moreover, newspaper articles, blog posts and personal diaries convey the same function as well. Most of the interpersonal bundles in the GLEAN corpus are stance features, which are found to be a distinctive feature of the soft knowledge fields (Hyland, 2008; Durrant, 2017). They convey a degree of certainty (15), possibility (16) or importance (17).

(15) ***In all likelihood****, after the referendum in July voters will choose 'yes ' or 'no' for further bailout air.*

(16) *I would like to say that* ***it is possible to*** *stabilize the Lari, as in summer we expect a lot of foreign currency to flow into Georgia.*

(17) ***It is important for*** us to know for sure that everything is going fine there.

The final functional participant-oriented bundle sub-type is the engagement feature, which addresses the reader directly (18, 19, 20). It explicitly marks the presence of the reader and acknowledges the dialogic dimension of narrative or research writing (Hyland, 2008: 19).

(18) *Follow me* ***to learn more about*** *the most common Eastern traditions in Georgia.*

(19) ***Our thoughts are with*** *the victims and those who are affected by the deadly floods in Tbilisi caused by the heavy rainfalls that hit the country over the past two days.*

(20) ***Our main goal*** *today is to put a huge political full stop to the Georgian Dream.*

5. Discussion and lexicographic implications

In the majority of bilingual dictionaries, lists of formulaic language items are presented as static and decontextualized lists, but it is hoped that by lemmatizing the most common lexical bundles as separate dictionary entries, students will be able to see the bundles at work in the texts they read and write. This is important because one of the pedagogical challenges of working with lexical bundles is a lack of definitions and illustrative sentences in the dictionaries for some students (Byrd & Coxhead, 2010).

Concerning the quantitative and qualitative findings discussed above, we will present a ready-to-use dictionary entry to ensure that this research can be directly applied to specialized bilingual lexicography.

We have created a sample entry for “as for the” on the Lexonomy platform (Mechura, 2017) (See Figure 2).

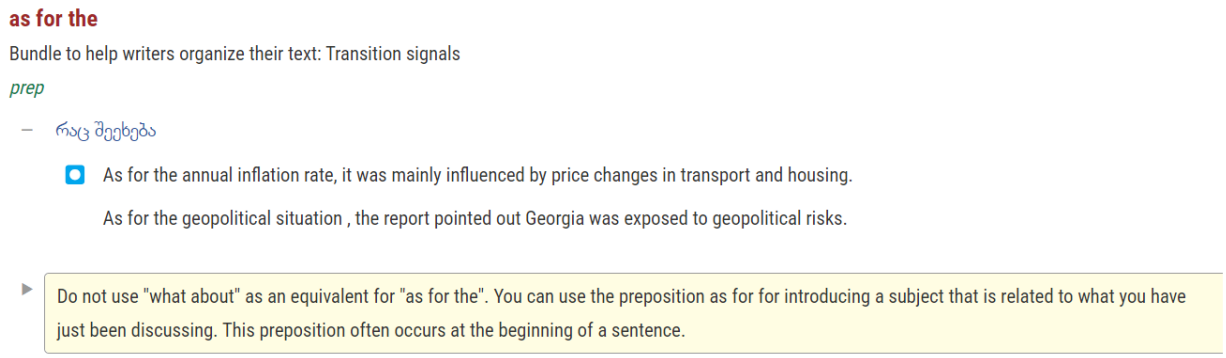

Figure 2 A sample entry for 'as for the' in a LEAD-style electronic dictionary

The microstructure of the lemma "as for the" is based on the learner corpus we have created, and it consists of a) the head phrase; b) the functional feature based on the functional taxonomy; c) the Georgian equivalent for the lexical bundle "as for the"; d) illustrative sentences retrieved from the Georgian Learner Corpus of English; e) the usage notes, or help boxes, which include warnings against some erroneous usage, e.g. false friends and calques. First of all, the head phrase is given as a lexical bundle fully and not dissected as separate lexical units. Due to the linguistic needs of the EFL students who want to improve their writing and compose coherent texts, we also decided to include the functional features of the bundles. In this case, a sample lexical bundle "as for the" represents a transition signal with the function or text organization. Furthermore, the illustrative sentences incorporated in the sample dictionary entry are taken from the GLEAN corpus as Georgian learners of English produce them and are authentic by nature. The last paradigm of the microstructure of the entry is the usage note element, where we warn the readers not to use "what about" as an equivalent for "as for the". Learner corpus analysis showed some erroneous usage of this lexical bundle in the declarative sentence.

6. Conclusion

Based on the GLEAN learner corpus we created, the paper has discussed and identified which lexical bundles (or lexical phrases) are most common in the academic prose produced by Georgian learners of English. We classified the functions of the most common 3-word or 4-word lexical bundles. We highlighted the value of adding learner corpus data (such as illustrative sentences, usage notes, and "help boxes") to learner dictionaries. At every stage of the corpus

development, the lexical bundle list, resources and methodological decisions were guided by lexicographic considerations. The results showed that the most frequent bundles in the Georgian Learner English corpus are the bundles serving the primary function of participant-oriented bundles, which serve the purpose of expressing attitudes.

The final product is a list of 42 lexical bundles covering academic writing in the GLEAN corpus disciplines (literature, linguistics, press, blog posts, etc.). It also discusses the lexicographic implications for further research. I hope this paper will spur more researchers and lexicographers to develop disciplinary lists of lexical bundles and incorporate their findings into bilingual English-Georgian dictionaries.

Acknowledgements

I want to express my deepest and most sincere gratitude to my PhD supervisor, Professor Manana Rusieshvili, for her infectious enthusiasm, encouragement, and intellectual perceptiveness. I am also indebted to Professor Sebastian Hoffmann of the Trier University in Germany for his unfailing expert guidance and support in compiling the learner corpus.

References:

Altenberg, B. (1998). *On the phraseology of spoken English: The evidence of recurrent word-combinations*. na.

Anthony, L. (2014). AntConc (Version 3.4. 3)[Computer Software]. Tokyo, Japan: Waseda University.

Bergenholtz, H., & Gouws, R. H. (2013). The Presentation of Word Formation in General Monolingual Dictionaries. Lexikos, 23, 59-76.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (2000). Longman grammar of spoken and written English.

Biber, D., Conrad, S. & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3), 371-405.

Byrd, P. & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL*, 5(5), 31-64.

Dirk, S. (2004). *Discourse markers across languages: A contrastive study of second-level discourse markers in native and non-native text with implications for general and pedagogic lexicography*. Routledge.

Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics*, 38(2), 165-193.

Erman, B. & Warren, B. (2000). The idiom principle and the open choice principle. *Text & Talk*, *20*(1), 29-62.

Gates, E. (1988). 'The treatment of multi-word lexemes in some current dictionaries of English', in Snell-Hornby (1986): 99–106.

Gledhill, C. J. (2000). *Collocations in science writing* (Vol. 22). Gunter Narr Verlag.

Gouws, R. H. (1991). Toward a lexicon-based lexicography. Dictionaries: Journal of the Dictionary Society of North America, 13(1), 75-90.

Granger, S. (2009). Prefabricated patterns in advanced efl writing: Collocations and formulae (oup, 1998).

Granger, S. & Paquot, M. (2009). Lexical verbs in academic discourse: A corpus-driven study of learner use. *Academic writing: At the interface of corpus and discourse*, 193-214.

Granger, S. (2018). Formulaic sequences 1 in learner corpora: Collocations and Lexical Bundles. *Understanding formulaic language*, 228-247.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International journal of applied linguistics*, 4(2), 237-258.

Hausmann, F. J. (1993): Was ist eigentlich Wortschatz? In: W. Börner & C. Vogel (Hrsg.): *Wortschatz und Wortschatzerwerb*. Bochum, 2-21.

Hoffmann, S., & Evert, S. (1996). BNCweb (CQP-edition). URL: http://bnc-web. lancs. ac. uk (Last accessed on March 28, 2024).

Hoffmann, S., Evert, S., Smith, N., Lee, D. & Berglund-Prytz, Y. (2008). Corpus linguistics with BNCweb-a practical guide.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes*, 27(1), 4-21.

Lewis, M. (1993) The Lexical Approach: The State of ELT and a Way Forward, Hove: Language Teaching Publications.

Makhatadze, M. (2023). Perfectly Perfect Adverbs: A Learner Corpus Study of the Amplifier Collocations by Georgian Learners of English.

Měchura, M. B. (2017, September). Introducing Lexonomy: an open-source dictionary writing and publishing system. In Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference (pp. 19-21).

Milton, J.,& Cheng, V. S. (2010, June). A toolkit to assist L2 learners become independent writers. In Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing processes and authoring aids (pp. 33-41).

Nesselhauf, N. (2005). Collocations in a learner corpus. *Collocations in a Learner Corpus*, 1-344.

North, S. (2005). Different values, different skills? A comparison of essay writing by students from arts and science backgrounds. *Studies in higher education*, 30(5), 517-533.

O'Flynn, J. (2022). Lexical bundles in the academic writing of the Arts and Humanities: from corpus to CALL. *Yearbook of Phraseology*, 13(1), 81-108.

O'Flynn, J. (2019). *Developing an Academic Collocation List for Arts and Humanities* (Doctoral dissertation, Master's Dissertation]. University of Warwick).

Simpson-Vlach, R. & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. Applied linguistics, 31(4), 487-512.

Sinclair, J. 2010 [2007]. Defining the Definiendum. In de Schryver, G.-M. (ed.) *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festchrift for Patrick Hanks*. Kampala: Menha Publishers, 37-47.

Sinclair, J. (1991). Corpus, concordance, collocation. (No Title).

Zgusta, L. (2010). *Manual of lexicography* (Vol. 39). Walter de Gruyter.

Appendix A.

Order	Lexical Bundle	GLEAN corpus Raw frequency	Frequency (per million word)
1.	As of today	1,209	123.658
2.	As well as the	1,173	119.976
3.	As a result of	864	88.371
4.	By the end of	794	81.212
5.	In addition to	627	64.131
6.	At the same time	595	60.858
7.	In this regard	390	39.890
8.	In the process of	384	39.276
9.	As for the	383	39.174
10.	In accordance with	383	39.174
11.	With the support of	358	36.617
12.	On the backdrop of	315	32.219
13.	In the wake of	300	30.684
14.	In response to the	203	20.763
15.	The need for	197	20.149
16.	In the case of	193	19.740
17.	On a daily basis	156	15.956
18.	In the context of	140	14.319
19.	At the beginning of	133	13.603
20.	In comparison to	118	12.069
21.	In a number of	110	11.251
22.	It is important for	102	10.433
23.	To learn more about	98	10.024
24.	In the middle of	97	9.921
25.	On the one hand	97	9.921
26.	In recognition of	91	9.308

27.	In honour of	83	8.489
28.	On the contrary	81	8.285
29.	In an attempt to	73	7.467
30.	I believe that the	53	5.421
31.	On the top of	47	4.807
32.	It is impossible to	45	4.603
33.	In the focus of	35	3.580
34.	At the heart of	33	3.375
35.	No matter how	31	3.171
36.	Our main goal	26	2.659
37.	Our thoughts are with	25	2.557
38.	Plays an important role	22	2.250
39.	I remain confident that	13	1.330
40.	I wish to remind	12	1.227
41.	Despite the fact that	3	0.339
42.	Also highlighted the fact	2	0.205

Author's email: marine_makhatadze@yahoo.com

Author's biographical data

Marine Makhatadze is a PhD student of Lexicography at Ivane Javakhishvili Tbilisi State University, Faculty of Humanities. She teaches at the Department of English Philology (the courses include Fundamentals of Lexicography and Learner Lexicography). She is interested in corpus linguistics, onomasiological dictionaries and is currently researching corpus-based learners' and native speakers' use of recurrent word combinations.